Rap Keb, behind the scene

Par Michaël Laforest

Au commencement de ce trimestre d'hiver, je suis tombé tout à fait par hasard sur un article du pudding, un site d'information qui use du journalisme de données pour proposer des articles tout à fait originaux. L'article de Matt Daniels dressait un portrait du rap aux États-Unis en basant son analyse sur le langage des différents artistes. Pour ce faire, il a extrait les paroles des chansons à partir de Genius, une bibliothèque de paroles à laquelle les internautes sont invités à participer. Pour chaque artiste, le journaliste retenait les 35 000 premiers mots prononcés par un artiste sur des albums studio. Ceux n'ayant pas atteint cette frontière étaient automatiquement rayés de son reportage. Ensuite, il en a fait un classement, il tentait de classer les rappeurs en fonction de l'étendue de leur vocabulaire tout en mettant de l'avant la décennie qui avait vu émerger chacun d'entre eux.

Lorsqu'on a proposé en classe de faire un reportage incluant un moissonnage et une analyse de données, cet article m'est tout de suite revenu à l'esprit. Considérant la popularité croissante du hip-hop au Québec, il me semblait tout à fait intéressant et original de tenter de l'analyser à l'aide de données, ce qui, jusqu'à aujourd'hui, n'a jamais été fait.

La première étape du projet a été de déterminer les artistes ayant eu une influence sur le hip-hop québécois, de ses débuts au Québec à aujourd'hui. La liste regroupait au départ un bassin intéressant et assez représentatif du rap québécois. Ensuite, il fallait récolter les données, et donc vérifier, pour chacun des artistes dans la liste, la quantité de paroles qu'il était possible de récupérer. Tout comme Matt Daniels, nous nous sommes tournés vers Genius pour récolter les données. Bien entendu, la filière québécoise du hip-hop est très sous-représentée, il était donc ardu de retrouver tous les artistes que nous avions préalablement identifiés, spécialement les plus anciens.

Extraction de données

Maintenant, il fallait s'atteler à trouver une manière de sauvegarder l'ensemble des paroles de chacun des artistes. J'ai été très heureux d'éviter la belle soupe et d'apprendre que Genius disposait d'un API. Après quelques recherches, j'ai trouvé un code plutôt simple me permettant d'extraire les données en plus d'explications sur ce qu'il était possible de faire en utilisant l'API de Genius. Je me suis donc créé un compte pour accéder à cet API et, après quelques expérimentations, j'ai réussi à extraire l'ensemble des paroles, et ce, pour les 52 artistes que nous avions identifiés et qui se retrouvaient dans la base de données.

Les données étaient ainsi compilées dans des fichiers .json, il y en avait 52 au total, un pour chacun des artistes. Il était maintenant temps de trouver une façon de séparer chacun des mots qui se trouvaient dans chacune des chansons. C'est ici que la fonction word_tokenize de nltk fut utile. En fouinant un peu dans le tutoriel et avec l'aide inestimable d'un certain professeur, il fut relativement facile de créer un fichier .csv dans lequel chaque ligne représentait un mot qui était lui-même associé à l'artiste, l'album, l'année et au nom de la chanson. Puis, en créant une boucle, il était possible de déterminer les nombres de mots différents qui se retrouvaient dans chacun des fichiers, et ainsi, nous avons réalisé notre premier objectif. La Constellation remportait du même coup la palme du collectif d'artistes le plus volubile dans l'histoire du hip-hop québécois, tel que déterminé par deux étudiants de l'UQAM.

Un défi supplémentaire

Au cours de nos recherches, nous en sommes également venus au constat que d'analyser la proportion d'anglais se retrouvant dans le vocabulaire de chaque artiste, ainsi que son évolution à travers le temps, serait une donnée intéressante à étudier. Puisque ce sujet soulève les passions de la société québécoise, il devenait incontournable. Je dois avouer humblement que, pour cette donnée, l'aide de Jean-Romain et Louis, nos deux collègues de Polytechnique, fut inestimable. La visualisation n'est peut-être pas à la hauteur de nos attentes, mais sans eux, mes cauchemars seraient encore remplis de séquences incompréhensibles de code et de boucles infinies.

Le plus gros problème de ce projet était de trouver un moyen de comparer tous les artistes sur un pied d'égalité. Nous avons d'abord tenté de catégoriser la volubilité par rapport aux albums plutôt que par rapport aux artistes, puis avons finalement opté pour un échantillon aléatoire de 3500 mots par artistes, ce qui, au final, permet tout autant une comparaison assez juste de chacun d'entre eux. Malgré la faible représentation du Québec dans la base de données de Genius et sans que nos données soient tout à fait exhaustives, elles offrent tout de même une certaine base pour pousser une certaine réflexion, ce qui était notre but initial.